TECNIACUSTICA
BARCELONA • 1996

Ponderación ML de parámetros en un sistema de reconocimiento de palabras basado en CDHMM

*A.J. Valverde, J. Hernando, A. Nogueiras**

Dpto. Teoría de la Señal y Comunicaciones. Universitat Politécnica de Catalunya
javier@gps.tsc.upc.es

ABSTRACT

Speech dynamic feature are routinely used in current speech recognition systems in combination with short-term (static) spectral features. The aim of this paper is to propose a method to automatically estimate the optimum ponderation of static and dynamic features in a speech recognition system. The recognition system considered in this paper is based on Continuous-Density Hidden Markov Modelling (CDHMM), widely used in speech recognition. Our approach consists basically in 1) adding two new parameters for each state of each model that weight both kinds of speech features, and 2) estimating those parameters by means of a Maximum Likelihood training. Experimental results in speaker independent digit recognition show an important increase of recognition accuracy.

INTRODUCCION

Las llamadas características dinámicas [1] se utilizan habitualmente en sistemas de reconocimiento del habla en combinación con las características estéticas. Dado que en su cálculo se utilizan varias tramas adyacentes, los parámetros que representan dichas características dinámicas modelan la evolución temporal del espectro de la Señal de voz. El uso de estos parámetros dinámicos reduce notablemente el número de errores en reconocimiento.

Aunque la mayoría de los sistemas de reconocimiento del habla existentes no poderan las características dinámicas con respecto a las estáticas, parece conveniente utilizar algún tipo de ponderación con el propósito de incrementar la tasa de reconocimiento del sistema. En los casos que se ha practicado la ponderación, ha sido ajustando las ponderaciones de forma manual [2], compensando las varianzas de ambos tipos de características [3], o proponiendo una fórmula de reestimación empírica para dichas ponderaciones [4].

El objetivo de este artículo es proponer un método para estimar de forma automática las ponderaciones óptimas de los parámetros estáticos y dinámicos en un sistema de reconocimiento del habla. El sistema de reconocimiento utilizado en el presente artículo está basado en los Modelos Ocultos de Markov de Densidad Continua (CDHMM) [5], ampliamente utilizados en el reconocimiento de voz. Nuestro trabajo consiste básicamente en 1) introducir dos nuevos parámetros para cada estado de cada modelo que ponderen ambos tipos de características del habla, y 2) estimar dichos parámetros de forma automática mediante un algoritmo de Máxima Verosimilitud.

PONDERACION DE CARACTERISTICAS DEL HABLA EN CDHMM

En los modelos ocultos de Markov de densidad continua, para un estado j del modelo i, la probabilidad de observar el vector de características O_t , $b_{ij}(O_t)$, se modela habitualmente mediante una mezcla de M funciones gaussianas multivariadas, es decir

$$b_{ij}(O_t) = \sum_{m=1}^{M} c_{ijm}\, N(O_t, \mu_{ijm}, U_{ijm}) \qquad (1)$$

donde N(.) es una función de densidad de probabilidad gaussiana con vector de medias μ_{ijk} y matriz de covarianza U_{ijk}, y c_{ijk} es el coeficiente de mezcla.

(1) * Este trabajo ha sido financiado por los proyectos TIC95-0884-C04-02 y TIC 95-1022-C05/03

Cuando se utilizan las características dinámicas, el vector O_t se forma habitualmente concatenando dichas características a las estáticas. Una aproximación alternativa consiste en considerar dos vectores separados O_t^1 y O_t^2, uno para cada tipo de característica, y asumir que ambos vectores son estadísticamente independientes, quedándonos en este caso

$$b_{ij}(O_t) = \prod_{s=1}^{2} \sum_{t=1}^{M} c^s_{ijm}\, N(O^s_t, \mu^s_{ijm}, U^s_{ijm}) \qquad (2)$$

donde N(.) es una función de densidad de probabilidad gaussiana con vector de medias μ^s_{ijk} y matriz de covarianzas U^s_{ijk}, y c^s_{ijk} es el coeficiente de mezcla.

Muchos sistemas de reconocimiento basados en CDHMM utilizan matrices de covarianzas diagonales con el propósito de aumentar la entrenabilidad de los modelos y reducir el cálculo computacional, siendo en este caso inmediato demostrar la igualdad de las expresiones (1) y (2). Sin embargo, la forma separada de (2) permite introducir de forma directa e intuitiva una ponderación muy simple sobre ambos tipos de vectores a través de unos pesos exponenciales $\{\omega^s_{ij}\}_{s=1,2}$ del siguiente modo

$$b_{ij}(O_t) = \prod_{s=1}^{2} \left[\sum_{t=1}^{M} c^s_{ijm}\, N(O^s_t, \mu^s_{ijm}, U^s_{ijm}) \right]^{\omega^s_{ij}} \qquad (3)$$

Esta nueva formulación ser la que utilizaremos y los dos nuevos parámetros para cada estado de cada modelo serán estimados mediante un algoritmo de Máxima Verosimilitud que pasaremos a detallar a continuación.

ESTIMACION DE MAXIMA VEROSIMILITUD DE LOS PARAMETROS DE PONDERACION

La aplicación del principio de Máxima Verosimilitud a la hora de maximizar $b_{ij}(O_t)$ requiere imponer algún tipo de restricción sobre las ponderaciones (sij si se quieren tener valores finitos de estos parámetros. Debido a las dificultades iniciales para encontrar una restricción simple y adecuada que nos proporcionase un algoritmo eficiente para obtener las ponderaciones, se propuso [6] un método para estimar automáticamente las ponderaciones óptimas de los parámetros estáticos y dinámicos mediante un algoritmo de entrenamiento discriminativo basado en el método GPD (Global Probabilistic Descent) [7]. Sin embargo, se continuó la búsqueda de una restricción adecuada sobre las ponderaciones que fuese consistente con la estimación por Máxima Verosimilitud, que es la convencional para el resto de parámetros del modelo. Finalmente, en el presente artículo mostramos cómo la introducción de una restricción simple sobre las ponderaciones de la forma

$$\sum_{s=1}^{2} (\omega^s_{ij})^m = K \qquad m \neq 1 \qquad (4)$$

proporciona un algoritmo de reestimación simple y eficiente, con el que se obtienen importantes mejoras en el reconocimiento para valores adecuados de K y m. Utilizando la función auxiliar de Baum en función de las probabilidades forward-backward y la técnica de multiplicadores de Lagrange, es fácil llegar a la siguiente fórmula de reestimación para las ponderaciones

$$\omega^s_{ij} = \left[K \cdot \frac{\sum_{r=1}^{R} \frac{1}{P_r} \left[\left(\sum_{t=1}^{T_r} \left(\alpha_{ij}(t) \cdot \beta_{ij}(t) \cdot \log b^s_{ij}(O^r_t) \right) \right)^{\frac{m}{m-1}} \right]}{\sum_{r=1}^{R} \frac{1}{P_r} \left[\sum_{s=1}^{2} \left(\sum_{t=1}^{T_r} \left(\alpha_{ij}(t) \cdot \beta_{ij}(t) \cdot \log b^s_{ij}(O^r_t) \right) \right)^{\frac{m}{m-1}} \right]} \right]^{\frac{1}{m}} \qquad (5)$$

donde $\alpha_{ij}(t)$ y $\beta_{ij}(t)$ son las probabilidades adelante y atrás, respectivamente, R es el número de pronunciaciones o secuencias independientes, T_r es el número de tramas de la secuencia r y P_r es la probabilidad de que el modelo entrenado genere la secuencia r.

RESULTADOS EXPERIMENTALES

Bases de Datos y Sistema de Reconocimiento

Las pruebas se han realizado sobre la base de datos TI [8] que consta de un gran número de secuencias de dígitos en inglés pronunciados en ausencia de ruido. La base de datos originalmente muestreada a 20 kHz se convirtió a 8 kHz. Para las pruebas se utilizaron sólo los locutores adultos, siendo éstos 112 de entrenamiento y 113 de test, y los dígitos aislados (one, two, ..., nine, zero, oh), con un total de 2464 secuencias para entrenamiento y 2486 para test.

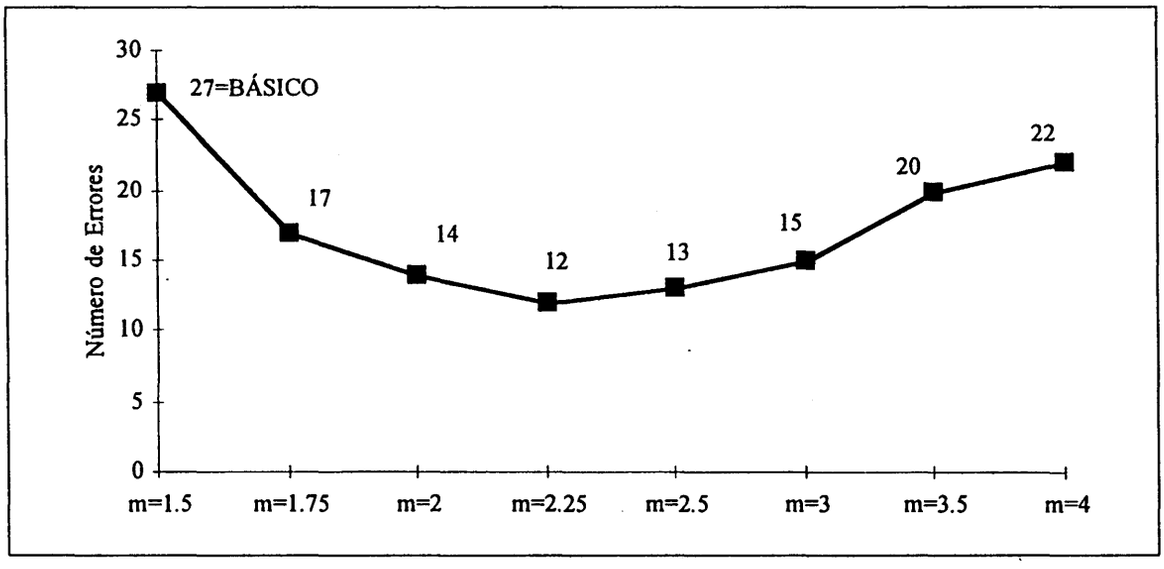
El sistema de reconocimiento utilizado es el HTK v1.5 [9], el cual está basado en modelos ocultos de Markov de densidad continua. Cada dígito se caracterizó por un modelo de Markov de 10 estados de izquierda a derecha sin saltos, una mezcla por estado con matriz de covarianza diagonal. El silencio se caracterizó mediante un modelo del mismo tipo, con pero con 5 estados.

La parametrización utilizada ha sido la LPC-cepstrum, formando un vector con 12 coeficientes cepstrales obtenidos a partir de un modelo de predicción lineal de orden 10 y la energía, y otro vector con los parámetros diferenciales correspondientes, delta-cepstrum y delta-energía [1]. El entrenamiento de los modelos consta de las siguientes etapas: inicialización mediante el algoritmo Segmental K-means, con segmentación manual previa, y una etapa con 5 iteraciones de reestimación mediante el algoritmo Baum-Welch. En la etapa de reconocimiento se utilizó el algoritmo de Viterbi clásico.

Resultados de Reconocimiento

A continuación se presentan los resultados obtenidos en reconocimiento de dígitos aislados. Tras estudios preliminares, se observó la poca sensibilidad del sistema a variaciones de la constante K. Dado que los valores de los pesos por defecto son la unidad y que se ha trabajado con dos vectores de información, se ha fijado el valor de K a 2.

El número total de errores de reconocimiento obtenido utilizando el sistema básico, es decir, $\omega_{ij}^{s} = 1$, ha sido de 27 sobre las 2486 secuencias de test. En la gráfica 1 se puede apreciar el número de errores obtenido al efectuar un barrido en el exponente m de la restricción (4) propuesta.



Gráfica 1. Número de errores al hacer un barrido del exponente m para K=2.

Tal y como se puede observar, excepto para m = 1.5 que se iguala el sistema básico, se mejora el reconocimiento para todos los valores de m estudiados, obteniéndose un mínimo de 12 errores para m = 2.25, con una reducción del 55.5 %. Para el caso especial de m = 2, correspondiente a la norma euclídea de l vector de pesos de un estado, se obtiene también un buen resultado de 14 errores. Se ha comprobado que para valores mayores a los presentados de m el número de errores se encuentra siempre por debajo del sistema básico.

Cabe destacar que, tras la reestimación, las ponderaciones correspondientes a los vectores estáticos estaban próximas y por debajo de la unidad, mientras que las correspondientes a los dinámicos estaban próximas y por encima de la unidad.

CONCLUSIONES

La inclusión de los parámetros de ponderación de las caracteristicas dinámicas de la Señal de voz frente a las estáticas y la estimación de los mismos mediante un algoritmo de Máxima verosimilitud proporcionan mejoras muy importantes en el reconocimiento de palabras aisladas mediante modelos ocultos de Markov de densidad continua.

Para aplicar un algoritmo de máxima verosimilitud se ha impuesto una restricción sobre los pesos. Esta restricción, necesaria para garantizar que el algoritmo proporcione una solución finita, consiste en fijar la norma módulo m del vector de pesos de cada modelo de cada estado a un valor constante K. Se ha comprobado que los valores de m y K [illegible] para la tarea emprendida. En particular, se han obtenido excelentes resultados para m = 2 (norma euclídea) y K = 2 (el valor que tendría esta constante por defecto, con pesos unitarios).

REFERENCIAS

[1] S. Furui, IEEE Trans. ASSP, vol. 34, pp. 52-59, 1986.

[2] K.F. Lee, de. Kluwer Academic Publishers, 1989.

[3] J.G. Wilpon, C.H. Lee, L.R. Rabiner, Proc. ICASSP-91, pp. 349-352, Toronto, 1991.

[4] C. Martín del Alamo, F.J. Caminero Gil, et.al., EUROSPEECH93, pp. 93-96, Madrid, 1995.

[5] L.R. Rabiner, B.H. Juang, IEEE ASSP Magazine, vol. 3, n 1, pp. 4-16, 1986.

[6] J. Hernando, J. Ayarte, E. Monte, Proc. EUROSPEECH95, pp. 105-108, Madrid, 1995.

[7] B.H. Juang, S. Katagiri, IEEE Trans. ASSP, vol. 40, n 12, pp. 3043-3054, 1992.

[8] R.G. Leonard, Proc. ICASSP84, pp. 42.11.1-4, 1984.

[9] HTK -Hidden Markov Model Toolkit v1.5, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc., 1993.